VIROLOGY 213, 169–178 (1995)

The Flavivirus 3′-Noncoding Region: Extensive Size Heterogeneity Independent of Evolutionary Relationships among Strains of Tick-Borne Encephalitis Virus

GERHARD WALLNER, CHRISTIAN W. MANDL,[1] CHRISTIAN KUNZ, and FRANZ X. HEINZ

Institute of Virology, University of Vienna, Vienna, Austria

Received June 2, 1995; accepted August 10, 1995

The sequences of the 3′-noncoding regions (NCR) of 12 strains of tick-borne encephalitis (TBE) virus were analyzed and found to vary in length from 350 to approximately 750 nucleotides. The size heterogeneity is restricted to a variable region following the stop codon, whereas the most 3′-terminal 350 nucleotides form a highly conserved core element containing several potentially important sequence motifs and secondary structure elements. A homoadenosine tract previously thought to form the 3′-terminus of some TBE virus strains was now shown to be an internal part of the variable region of certain strains. The strains included in this study were isolated from both humans and ticks over a time period of more than 40 years at various locations throughout the entire endemic area of TBE virus, but there was no correlation between these parameters and the observed lengths of the 3′-NCRs. Identity data calculated from common 3′-NCR sequences and also from short sections of the open reading frame indicated that coding and noncoding sequences were linked during evolution, but the lengths of the 3′-NCRs were independent of these relationships. These observations together with detailed analyses and alignments of the sequences suggest that the variable region was originally acquired through duplication and recombination events, but — much more recently during evolution — various portions of this region were lost again, resulting in the now observed heterogeneous 3′-NCRs.

INTRODUCTION

The three genera of the family Flaviviridae, i.e., *Flavivirus, Pestivirus,* and hepatitis C virus, comprise enveloped viruses with positive-stranded RNA genomes that share a common genomic organization (Wengler *et al.,* 1995). All viral proteins are encoded within a single long open reading frame in the gene order 5′–capsid protein–envelope proteins–nonstructural proteins–3′. This open reading frame is flanked by relatively short noncoding regions (NCR) that are believed to contain important control elements of various viral functions including replication, translation, and packaging of the genome.

The genus *Flavivirus* consists of approximately 70 members, which are mostly arthropod-borne viruses, and they include a number of important human pathogenic disease agents (Calisher *et al.,* 1989; Westaway *et al.,* 1985). The term "flavivirus" in this communication shall stand exclusively for members of the genus rather than for all three genera of the flavivirus family. With respect to the arthropod vectors flaviviruses can be divided into two major groups, namely, mosquito-borne flaviviruses (e.g., yellow fever virus, dengue viruses, and Japanese encephalitis virus) and tick-borne flaviviruses. The most prominent member of the latter group is tick-borne encephalitis (TBE) virus, a disease agent endemic in many parts of Europe and Asia. At least two subtypes of TBE virus can be serologically distinguished, which are referred to as the European and the Far Eastern subtypes, respectively (reviewed by Heinz, 1986).

In the flavivirus genome, the 3′-NCR usually comprises between 400 and 600 nucleotides and thus is longer than the 5′-NCR. Computer analyses and sequence comparisons revealed a number of notable elements within the flaviviral 3′-NCR, for which functional importance has been implied (Hahn *et al.,* 1987). The 3′-terminal approximately 90 nucleotides are predicted by computer analyses to fold into a characteristic secondary structure, which is largely conserved between flaviviral 3′-terminal sequences in spite of considerable primary sequence differences (Rice *et al.,* 1985; Brinton *et al.,* 1986; Wengler and Castle, 1986; Mandl *et al.,* 1993). Detailed comparative sequence analyses (Mandl *et al.,* 1993) and some direct experimental evidence (Brinton *et al.,* 1986; Hahn *et al.,* 1987) support the existence of this structure *in vivo.* In addition to this conserved secondary structure a few potentially important primary sequence elements within the 3′-NCR were discovered on the basis of their significant conservation among flaviviruses: (i) the 3′-terminal two nucleotides are generally -CU-3′ (complementary to the 5′-terminal 2 nucleotides 5′-AG-), (ii) an

Sequence data in this article have been deposited with GenBank under Accession Nos. U27490 (132), U27491 (263), U27492 (Aina), U27493 (Crimea), U27494 (Ljubljana I), U27495 (Neudoerfl, complete genome), and U27496 (RK1424).

[1] To whom correspondence and reprint requests should be addressed at: Institute of Virology, Kinderspitalgasse 15, A-1095 Vienna, Austria. Fax: 43-1-406 21 61.

0042-6822/95 $12.00

absolutely conserved pentanucleotide motif 5′-CACAG-3′ is present within an unpaired loop of the 3′-terminal secondary structure (Mandl *et al.,* 1993), and (iii) there are a pair of 18- to 20-nucleotide-long sequence motifs preceding the 3′-terminal secondary structure which were termed CS1 and CS2 (Hahn *et al.,* 1987). These conserved boxes, however, are present only in mosquito-borne flaviviruses, but were not identified in the 3′-NCRs of tick-borne flaviviruses sequenced to date, i.e., TBE virus (Mandl *et al.,* 1991b) and Powassan virus (Mandl *et al.,* 1993). Functional activities that may be connected with both the secondary structure and the conserved sequence motifs include recognition of cellular or viral replication or translation factors or packaging of the genomic RNA into the viral capsid. Moreover, the 3′-terminal secondary structure very likely is needed to stabilize the RNA genome which — different from most other cellular or viral positive-stranded RNA molecules — lacks a 3′-terminal poly(A) tail. Furthermore, part of the conserved CS1 box was proposed to function as a cyclization sequence (Hahn *et al.,* 1987), inducing the formation of a so far hypothetical panhandle structure during replication in analogy to the situation with alphaviruses (reviewed by Strauss and Strauss, 1986). For tick-borne flaviviruses, an alternative potential cyclization sequence was localized within the 3′-terminal secondary structure (Mandl *et al.,* 1993).

The analysis of strains of TBE virus added a surprising new aspect to the structure of the flaviviral 3′-NCR. Two closely related strains both belonging to the European subtype of TBE virus were found to contain quite different 3′-NCRs (Mandl *et al.,* 1991b). Strain Hypr (isolated from human brain in Czechoslovakia in 1956) exhibited a 3′-NCR matching in length (461 nucleotides) and in its organization the usual flavivirus pattern described above. In contrast, the 3′-NCR of strain Neudoerfl (isolated from a tick in Austria in 1971) appeared to be only 130 nucleotides long and terminate with a poly(A) tail rather than the characteristic secondary structure of flaviviruses. Simultaneously, it was reported that the 3′-NCRs of hepatitis C viruses — reminiscent of TBE virus strain Neudoerfl — were rather short, variable sequences that may carry poly(A) or poly(U) tails (reviewed by van Doorn, 1994), whereas the 3′-termini of pestiviruses would rather resemble the other flaviviruses, ending with a conserved secondary structure and lacking any homopolymeric tail (Deng and Brock, 1993). In the light of these observations and considering the functional importance of the 3′-NCR we investigated the puzzling sequence heterogeneity of TBE virus more extensively. In this communication we report on the sequence analysis of the 3′-terminal genomic regions of 12 strains of TBE virus belonging to both the European and the Far Eastern subtypes. Our data illustrate a high variability within part of the 3′-NCR, including the existence of a poly(A) tract in some strains. These poly(A) tracts, however, are shown to be internal elements of the 3′-NCR, whereas the most 3′-terminal section of all TBE virus 3′-NCRs forms a highly conserved secondary structure. Potentially important sequence elements were identified and a common organizational scheme for all the TBE virus 3′-NCRs was derived. On the basis of these data we finally propose a hypothesis on the evolution of the extensive 3′-NCR sequence heterogeneity.

MATERIALS AND METHODS

Virus strains

Table 1 lists the TBE virus strains that were analyzed in this study. Strains from various parts of the endemic area of TBE virus, isolated over a time period of more than 40 years both from humans and ticks, were included. We gratefully acknowledge those colleagues who kindly provided us with some of these strains (Table 1).

Preparation of viral RNA and specific 3′-NCR cDNA

For the preparation of viral RNA from these strains, viruses were grown in BHK 21 cells. Supernatants (30 ml) from 77-cm^2 tissue culture flasks were harvested 48 hr postinfection. Cell debris was removed by a 30 min/10,000 rpm centrifugation in a Sorvall SS34 rotor and subsequently virus was pelleted by ultracentrifugation (30 min/44,000 rpm in a Beckman Ti45 rotor). The pellet was resuspended in 500 μl TAA buffer (0.05 *M* triethanolamine, pH 8.0, 0.1 *M* NaCl) supplemented with 0.1% BSA and subjected to the proteinase K/phenol/chloroform RNA extraction procedure described previously (Mandl *et al.,* 1988). After precipitation of the RNA the total yield from one such RNA preparation was put into a single cDNA synthesis reaction using the Boehringer-Mannheim cDNA synthesis kit and a synthetic oligonucleotide primer (Medprobe, Norway) complementary to the 18 3′-terminal nucleotides of TBE virus strain Hypr (Mandl *et al.,* 1991b). The polymerase chain reaction (PCR) was then employed to amplify various fragments corresponding to the 3′-NCR regions and the adjacent sections of the NS5 coding regions using conditions described elsewhere (Mandl *et al.,* 1989). The sequences of the 20-nucleotide-long PCR primers that were used in the original experiments were derived from the published NS5 coding region of strain Neudoerfl (Mandl *et al.,* 1989) for the positive-sense primers and the 3′-terminal sequence of strain Hypr (Mandl *et al.,* 1991b) for the antisense primers, respectively.

Sequence determination and computer analysis

PCR fragments were purified by phenol and chlorophorm extractions and then sequenced directly on both strands without intermediate cloning. Sequencing was performed on an ABI 373A automated sequencer using

TABLE 1

TBE Virus Strains

Strain designation	Source of isolation	Year of isolation	Geographic origin	Passage history[a]
Neudoerfl	*Ixodes ricinus*	1971	Austria, Burgenland	3
Alsace[b]	*Ixodes ricinus*	1975	France, Alsace	Unknown
263[c]	*Ixodes ricinus*	1987	Czechia, South Bohemia	3
Crimea[b]	*Ixodes ricinus*	1987	Ukraine, Crimea	4
RK1424[b]	*Ixodes persulcatus*	1977	Latvia	22
132[b]	*Ixodes persulcatus*	1979	Russia, Vladiwostok	Unknown
Absettarov[b]	Human (blood)	1951	Russia, St. Petersburg	Unknown
Hypr[d]	Human (blood)	1953	Czechia, Moravia	Unknown
Scharl	Human (brain)	1956	Austria, Lower Austria	8
Aina[b]	Human (blood)	1963	Russia, Irkutsk	Unknown
Golobokov[b]	Human (brain)	1981	Russia, Chabarovsk	3 or 4
Ljubljana I[e]	Human (blood)	1992	Slovenia, Ljubljana	4
Ljubljana II[e]	Human (blood)	1992	Slovenia, Ljubljana	3

[a] Number of passages by ic inoculation of suckling baby mice. For some of the strains the exact passage history is unknown, but probably consists of numerous passages.

[b] Kindly provided by Dr. M. Vorobyova, Tarasevich Institute, Moscow, Russia.

[c] Kindly provided by Dr. J. Kopecký, Institute of Parasitology, České Budějovice, Czech Republic.

[d] The nucleotide sequence of the 3′-NCR of strain Hypr has been elucidated previously (Mandl *et al.,* 1991b).

[e] Kindly provided by Dr. T. Avšič-Županc, Institute of Microbiology, Ljubljana, Slovenia. Strain Ljubljana II was isolated from a person who developed TBE after working with strain Ljubljana I in the laboratory (Avšič-Županc *et al.,* 1995).

fluorescent dye-labeled dideoxynucleotide terminators (Perkin–Elmer Corp., Applied Biosystems Division, Weiterstadt, Germany). After preliminary sequence information had been obtained, further strain-specific oligonucleotide primers were synthesized and employed for additional PCR and sequencing reactions. Sequence manipulations and comparisons were performed with the help of the Beckman Microgenie software package (version 4.0). Dendrograms based on multiple sequence alignments were drawn by use of the program CLUSTAL (Higgins and Sharp, 1988) contained in the *PC*Gene software package (IntelliGenetics, Geel, Belgium).

RESULTS AND DISCUSSION

Sequence analysis

The 3′-termini of TBE virus strains Hypr and Neudoerfl (Mandl *et al.,* 1991b) had previously been analyzed by ligation of the RNA 3′- and 5′-ends (after removal of the CAP structure), and cDNA synthesis across the joined termini followed by an amplification of this region by PCR (Mandl *et al.,* 1991a). In the approach used in this study we took advantage of the known sequence data and utilized a cDNA primer and subsequently a somewhat longer PCR primer complementary to the 3′-end of strain Hypr and a second PCR primer corresponding to a sequence selected from the 3′-end of the relatively conserved NS5 coding region. (The NS5 gene is located at the 3′-end of the long open reading frame.) Based on the available data it was assumed that this approach would yield PCR fragments in the cases of strain Hypr and other strains alike, but would fail to produce fragments representing the 3′-NCRs of strain Neudoerfl and other strains of its kind. To our surprise we obtained specific fragments for all of the virus strains listed in Table 1, including strain Neudoerfl. The sizes of these fragments, however, varied considerably. This heterogeneity is illustrated by Fig. 1, which shows an agarose gel analysis of PCR fragments spanning the genomes from nucleotide 9940 to the 3′-termini obtained from each strain by use of the same pair of PCR primers. As can be seen from the figure, the PCR fragment synthesized

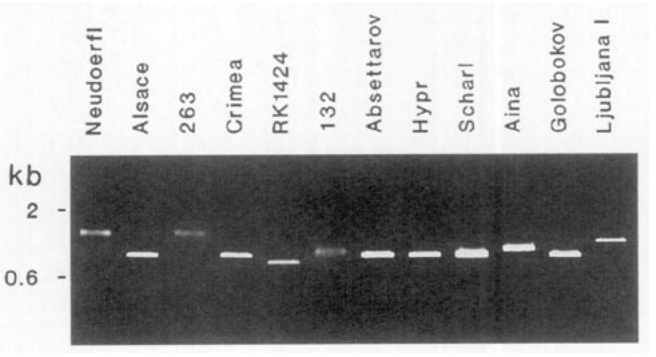


FIG. 1. PCR fragments spanning the 3′-NCRs of TBE virus strains analyzed on an ethidium bromide-stained agarose gel. The same cDNA and PCR primers were employed for each strain, i.e., cDNA primer H01 (5′-AGCGGGTGTTTTTCCGAG-3′) and PCR primer H02 (5′-AGCGGGTGTTTTTCCGAGTC-3′), which are complementary to the 18 and 20 3′-terminal nucleotides of TBE virus strain Hypr, respectively, and PCR primer S01 (5′-GGGCAGATGTGGCTGCTGAG-3′), which corresponds to nucleotides 9940 to 9959 of the TBE virus strain Neudoerfl, a sequence conserved among all of the strains analyzed.

from the cDNA of strain Hypr is approximately 900 bp long, in agreement with the published sequence data. Other strains of TBE virus yielded fragments as short as approximately 800 bp (strain RK1424) or as long as approximately 1300 bp (e.g., strain Neudoerfl). Interestingly, for each virus strain only a single fragment size was amplified, suggesting that the virus preparations of each individual strain were homogeneous with respect to their 3′-NCRs. However, the bands obtained for strains Neudoerfl, 263, and 132 appeared to be rather diffuse compared to the other bands.

The nucleotide sequences of the 3′-NCRs and part of the adjacent NS5 coding regions were consequently determined by direct sequence analysis of these PCR fragments. Figure 2 shows the exact genomic sequences of TBE virus strains Neudoerfl, 263, Ljubljana I, 132, Crimea, Aina, and RK1424 from position 10,000 (numbering according to the sequence of strain Neudoerfl) to the 3′-termini and also includes the previously published sequence of TBE virus strain Hypr. The sequences are aligned for maximum identity and a number of particular sequence elements that will be discussed below are marked on top of the nucleotide sequences. A schematic drawing of the structural organizations of these 3′-NCRs is shown in Fig. 3. The 3′-NCRs of the remaining strains listed in Table 1 were only partially analyzed to determine their structural organizations. These data indicated that the 3′-NCR structures of the strains Alsace, Absettarov, Golobokov, and Scharl were very similar to those determined for strains Hypr and Crimea. The 3′-NCR of strain Ljubljana II was found to be identical to the Ljubljana I strain sequence.

Structural organization of the 3′-NCR

Figures 2 and 3 illustrate the size variability of the 3′-NCR regions, the longest being those found in the strains Neudoerfl and 263, and the shortest being that of strain RK 1424, the 3′-NCR region of which is only 350 nucleotides long. Thus, the 3′-NCR regions of TBE virus can be much longer as well as considerably shorter than that of any other flavivirus analyzed so far. Furthermore, Figs. 2 and 3 make it clear that the variability is restricted only to the region immediately following the stop codon, whereas the most 3′-terminal 325 nucleotides are present in all of the strains and share a high degree of sequence identity. In fact, the sequence identity in this region exceeds that observed in the sections of the NS5 coding region included in our analysis (Fig. 2). This high degree of conservation suggests a functional importance of these 325 most 3′-terminal nucleotides, which in fact may represent a "core element" of the 3′-NCR and will further be referred to by this designation.

The core element

The core element of the 3′-NCR, in principle, appears to be sufficient for a viable virus genome as exemplified by strain RK 1424. The RNA sequence is nonrepetitive throughout the core element and includes a number of interesting sequence motifs and structures which are described in the following paragraphs.

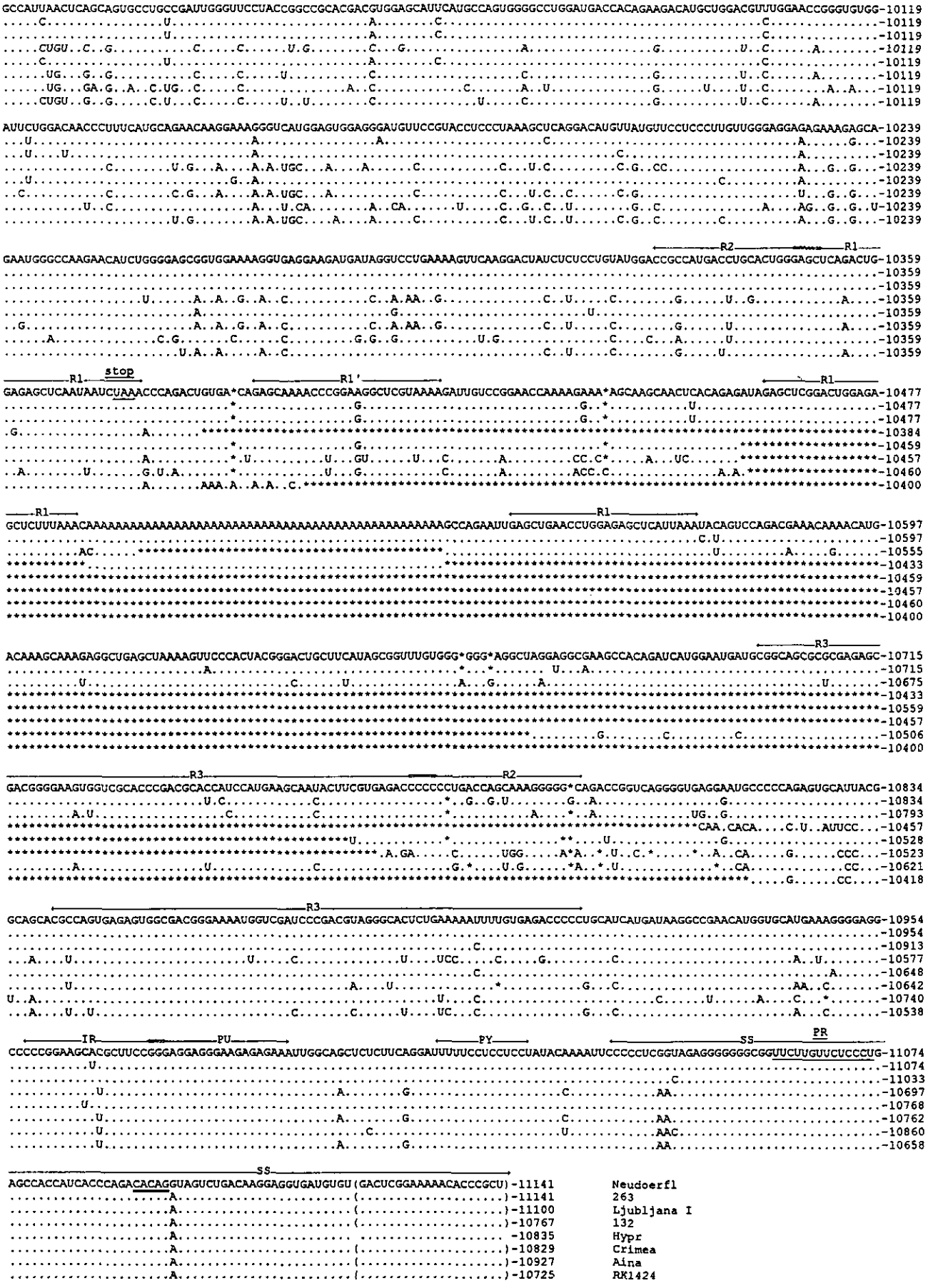
The 3′-terminal approximately 100 nucleotides consist of a sequence capable of folding into the secondary structure (ss in Figs. 2 and 3) that has been described before for tick-borne flaviviruses and which closely resembles that found for mosquito-borne flaviviruses. Nucleotide differences among strains in this region are present only in positions that are predicted not to base-pair in the structure and thus would not affect its conformation. The flavivirus-conserved CACAG box is contained in all the strains and is marked in Fig. 2. Our experimental approach implied that the original sequence of the most 3′-terminal 20 nucleotides could not be determined. The high degree of sequence conservation elsewhere within the 3′-terminal secondary structure, however, makes it likely that the 3′-terminal 20 nucleotides of these strains are very similar or even identical to the previously determined sequence of strain Hypr.

A short, but conserved inverted repeat structure is located somewhat more than 150 nucleotides away from the 3′-terminus (IR in Figs. 2 and 3). Due to its high GC content a relatively high thermal stability ($\Delta G = -19$ kcal) can be calculated, suggesting that this sequence may actually form a hairpin structure *in vivo*.

There are three distinguishable sequence elements located downstream of this hairpin structure which are conserved among all the TBE virus strains:

Adjacent to the IR sequence and overlapping by three nucleotides, there is a 20-nucleotide-long stretch of purine residues. This box (PU; see Figs. 2 and 3) is the significantly longest homopurine sequence element within the entire genome of TBE virus strain Neudoerfl apart from the poly(A) tract (see below). The second longest homopurine sequence within the TBE virus genome is only 16 residues long and not conserved among

FIG. 2. Nucleotide sequences from position 10,000 (numbering according to strain Neudoerfl) to the 3′-termini of TBE virus strains (from top to bottom) Neudoerfl, 263, Ljubljana I, 132, Hypr, Crimea, Aina, and RK1424. Nucleotides identical to the sequence of strain Neudoerfl are depicted as dots, and gaps introduced for the alignment are shown by asterisks. Individual sequence numbers are shown to the right. The most 3′-terminal 20 nucleotides are shown in parentheses to indicate that they were taken from the strain Hypr sequence. The length of the poly(A) tract is arbitrarily chosen to be 49 residues. Individual sequence elements are emphasized on top of the sequences: R1, R2, and R3, imperfect direct repeats; stop, first stop codon terminating the long open reading frame; IR, inverted repeat; PU, homopurine box; PY, homopyrimidine box; SS, 3′-terminal secondary structure; PR, pyrimidine-rich box; and the flavivirus-conserved CACAG box is emphasized by double underlining.


Neudoerfl
263
Ljubljana I
132
Hypr
Crimea
Aina
RK1424

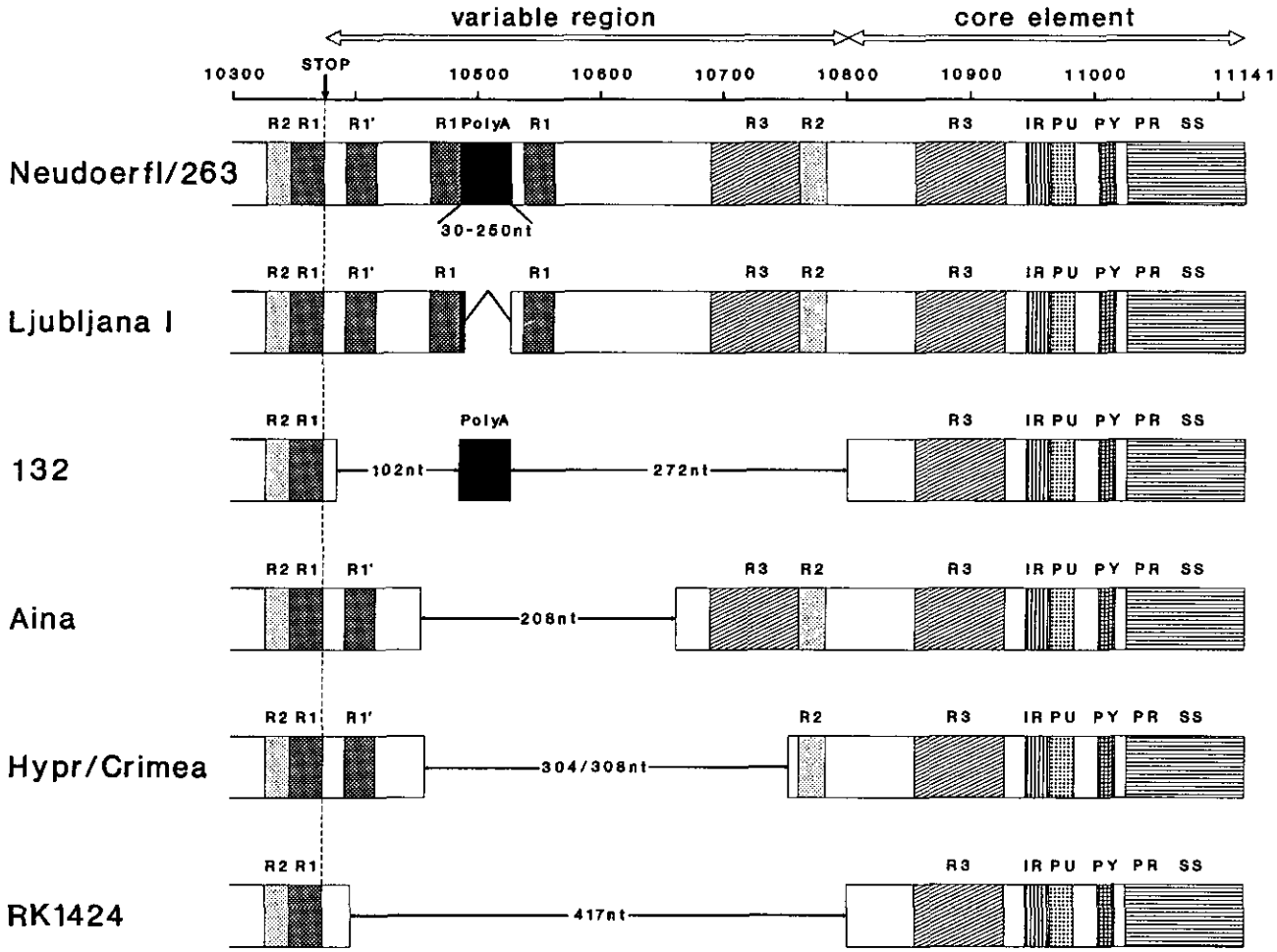


FIG. 3. Schematic drawing of the 3′-NCR structures of TBE virus strains. For the alignment of sequences only single gaps (two gaps in the case of strain 132) had to be introduced, the lengths of which are indicated. The 3′-NCR structures of strains Golobokov, Scharl, Alsace, and Absettarov are very similar to those shown for strains Hypr and Crimea, the corresponding gap sizes being 306, 314, 315, and 317 nucleotides, respectively. The 3′-NCR structure of strain Ljubljana II is identical with that of Ljubljana I. The 3′-NCRs are divided into a variable region and a core element as indicated at the top. Numbering corresponds to the strain Neudoerfl sequence. For the designations of individual sequence elements, please refer to the legend to Fig. 2.

strains (not shown). In addition to its length, an importance of this PU box is further suggested by the observation that an almost completely conserved homopurine box differing from the TBE virus element by only one A to G transition is also present at the same position within the genome of Powassan virus, which is a more distantly related tick-borne flavivirus (Mandl *et al.*, 1993).

A few nucleotides further downstream, there is a homopyrimidine sequence element (PY in Figs. 2 and 3), which, although consisting of only 14 residues, also represents the longest structure of its kind within the TBE virus genome. Its homologous counterpart in the Powassan virus genome, however, is interrupted by one purine residue.

In addition to the PU and PY boxes, another pyrimidine-rich (consisting of 13 pyrimidines and one purine) sequence motif (PR in Figs. 2 and 3) is located 30 nucleotides downstream of the PY box as a part of the 3′-terminal secondary structure. As can be seen in Fig. 2 this PR box is conserved among the TBE virus strains. A subsequence of PR is identical with the previously proposed potential cyclization element of tick-borne flaviviruses (Mandl *et al.*, 1993).

Although functional roles of these boxes remain hypothetical, it seems noteworthy that homopyrimidine and homopurine stretches are frequently found to be involved in regulatory processes of translation (Avni *et al.*, 1994; Behe, 1995), RNA stability (Czyzykrzeska *et al.*, 1994), or — e.g., in the case of retroviruses — replication (Hungnes *et al.*, 1992). For mosquito-borne flaviviruses, conserved sequence motifs (termed CS1 and CS2) are located at approximately the same genomic positions as the PU and PY boxes, i.e., a few nucleotides upstream of the 3′-terminal secondary structure (Hahn *et al.*, 1987). CS1 and CS2 are thought to be functionally important for the replication of mosquito-borne flaviviruses and a

subsequence of the CS1 box has been proposed to be a potential cyclization sequence. Although there is no significant sequence homology between the tick-borne elements PU, PY, and PR and the mosquito-borne CS1 and CS2 boxes, it is tempting to speculate that these sequence motifs might mediate analogous functional activities for the two groups of flaviviruses.

The variable region

The remaining part of the 3′-NCRs in between the stop codon and the core element is highly variable and accounts for all of the observed size differences between the TBE virus 3′-NCRs. Major sections of this hypervariable domain consist of imperfect direct repeat copies of sequences present either in the core element or in the adjacent NS5 coding region. The longest and most significant repeats are depicted as R1 and R2, which had already been recognized previously (Mandl *et al.*, 1991b), and R3 in Figs. 2 and 3. The most unusual building block of this domain is a homoadenosine sequence element. The presence of a poly(A) tract in the genome of prototypic strain Neudoerfl had already been discovered previously (Mandl *et al.*, 1991b) and was then believed to represent the 3′-terminus of the genome. The presence of a homoadenosine structure was now confirmed for this strain and in addition was recognized for other strains of TBE virus (strains 263 and 132), but it is shown to be an internal part of these 3′-NCRs rather than forming the 3′-terminus. It is interesting to note that the poly(A) tract is flanked by identical sequences in the cases of strains Neudoerfl and 263, but these flanking sequences are missing on both sides of the strain 132 poly(A) tract, which therefore is surrounded by a completely different sequence environment. On the other hand, the Ljubljana strains contain all of the sequence elements present in strains Neudoerfl and 263, but exhibit only an oligo(A) sequence, i.e., $(A)_4$-C-$(A)_6$, at the homologous position.

The actual length or range of lengths of the poly(A) tract in viable TBE virus genomes remains uncertain. In Fig. 2 it was depicted to be 49 residues long, because it was recently observed that infectious TBE virus genomes can be generated with this number of A residues (manuscript in preparation). Previous analyses had yielded poly(A) sizes ranging from 20 to more than 200 residues. As mentioned above, PCR-derived fragments spanning the poly(A)-containing regions were heterogeneous in size — they migrated as diffuse bands on agarose gels (Fig. 1) — but it is unknown how much of this heterogeneity was artificially caused by stuttering of the reverse transcriptase and/or the *Taq* polymerase and how much originated from an *in vivo* size heterogeneity. No heterogeneity was observed for the oligo(A) structure, i.e., $(A)_4$-C-$(A)_6$, in the two Ljubljana strains.

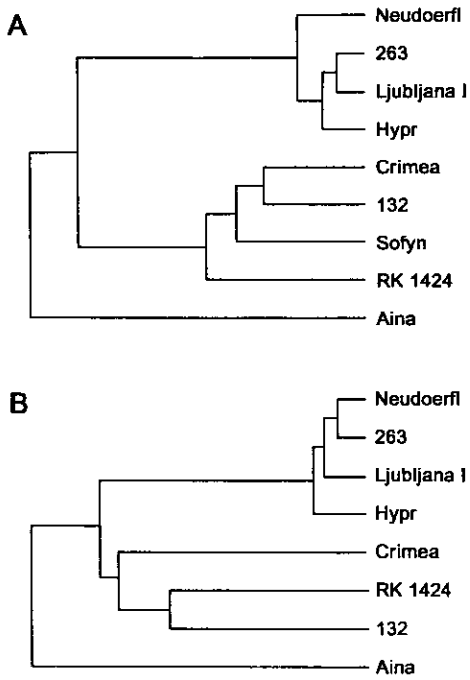


FIG. 4. Dendrograms reflecting the evolutionary relationships of TBE virus strains based on RNA sequence identities calculated (A) for 375 nucleotides of the NS5 coding region (from position 10,000 to the stop codon) and (B) for the 325 nucleotides of the core element of the 3′-NCR. The dendrograms were constructed by multiple sequence alignments using the program CLUSTAL with the following parameter settings: K-tuple value, 4; gap penalty, 12; filtering level, 5; window size, 20; open gap cost, 10; unit gap cost, 10; transitions weighted twice as likely as transversions.

Evolution of the 3′-NCR heterogeneity

How may the heterogeneity of the TBE virus 3′-NCR have evolved? To address this question we first wondered whether the occurrence of particular 3′-NCR structures correlated with any other strain-specific parameter, such as the year, geographic origin, or source of its isolation. Comparison of the schematic shown in Fig. 3 with the information listed in Table 1 indicates no connection between any of these parameters and the occurrence of a poly(A) structure or the length of the 3′-NCR.

Subsequently, we wanted to establish the sequence relationships among the strains independent of their 3′-NCRs. We took advantage of the partial NS5 coding sequences that were elucidated in our analysis (Fig. 2) and constructed an evolutionary tree based on the sequence identity data calculated from this 375-nucleotide-long section (Fig. 4A). The corresponding section of the published sequence of the Far Eastern subtype prototypic strain Sofyn (Pletnev *et al.*, 1990) was also included in this analysis. The dendrogram clearly demonstrates the close relationship of strains 263, Ljubljana I, and Hypr with Western subtype prototypic strain Neudoerfl, while strains 132, Crimea, and RK1424 are obviously more closely related with the Far Eastern subtype prototypic strain Sofyn. Strain Aina according to this analysis seems

more distantly related to both subtypic groups. These results correspond very well with the geographic origins of these strains (compare Table 1) and thus are in good agreement with previously reported serological data (Holzmann *et al.,* 1992) and RNA–DNA hybridization studies (Shamanin *et al.,* 1990). However, the structures of the 3′-NCRs do not match with this evolutionary tree: For example, strains Hypr and Neudoerfl are closely related Western subtypic strains, but have quite different 3′-NCR lengths, whereas the more distantly related Far Eastern strain Crimea shares an almost identical organization of its 3′-NCR with strain Hypr.

Since 3′-NCR lengths and evolutionary relationships derived from coding sequences thus were found not to match, we wondered whether 3′-NCRs and coding sequences evolved together or whether perhaps the different 3′-NCRs were acquired independently by recombination. To address this question we constructed another evolutionary tree, which now was based on the comparisons of the core elements of the 3′-NCRs, i.e., the 3′-terminal 325 nucleotides. This dendrogram turns out to be almost identical to the one constructed using the partial NS5 coding sequences (Fig. 4B). Similar results are obtained from the sequence comparisons of those sections of the variable regions of the 3′-NCRs that are shared by a subset of the strains. For example, the sequences from positions 10,766 to 10,816 of strains Hypr, Ljubljana I, 263, and Neudoerfl exhibit sequence identities among each other ranging from 86 to 92%, whereas the corresponding section of strain Crimea is clearly more distantly related, sharing only between 62 and 65% sequence identity with the sequences of the four Western subtypic strains. These data allow the conclusion that the coding sequences and the 3′-NCR sequences evolved together, but the formation of the size heterogeneity including the acquisition of a poly(A) structure occurred independently of this evolutionary process.

The generation of the different 3′-NCR lengths could, in principle, be explained either by sequence insertions into a small ancestral 3′-NCR or by deletions from a large ancestral 3′-NCR. In the cases of the TBE virus strains analyzed in this study we prefer to believe that these strains all descended from a common ancestor containing a 3′-NCR that was at least as long as those found in strains 263 and Neudoerfl. This hypothesis is based on the following considerations:

The variable regions of the various TBE virus strains, as shown in Fig. 3, can all be aligned smoothly by introducing only single gaps of variable lengths into each sequence (with the exception of the strain 132 sequence, where two gaps on both sides of the poly(A) tract had to be introduced for the alignment). None of the 3′-NCRs contains any sequence element that is not present also in the longest 3′-NCRs of strains Neudoerfl and 263. If, therefore, these different structures were to be formed by independent insertion events, one would have to assume very specific mechanisms for the generation of these insertions, in order to always end up with perfectly aligning sequences. Independently occurring deletion events from a common long ancestor 3′-NCR would be a much easier mechanism for generating these particular structures. The fact that sequence identities calculated for subsequences of the variable 3′-NCR section correspond to the evolutionary relationships derived from other parts of the genome (see the example given above) is in good agreement with the hypothesis that all of these sequences evolved from a common ancestor with a long 3′-NCR.

We further speculate that this common strain 263/Neudoerfl-like ancestor itself arose from an evolutionarily older virus with a smaller 3′-NCR, which in fact may have consisted only of a 3′-NCR core element. This assumption is supported by the observation that major parts of the variable region consist of direct repeats of sequences present in the core element or the adjacent NS5 coding region, suggesting that these sequences were acquired by duplication events.

Other parts of the variable region which exhibit no recognizable relationship with other genomic flavivirus sequences may have originally been acquired by recombination. Duplications, recombinational insertions, and deletions caused by template slippage or switching of the viral RNA polymerases have been described for several other RNA viruses (Meyers *et al.,* 1989; Dominguez *et al.,* 1990; Santagati *et al.,* 1994). Our data now provide evidence for similar mechanisms occurring in a flavivirus.

The origin of the poly(A) remains especially puzzling. Perhaps stuttering of the viral RNA polymerase caused a gradual increase in length of an originally shorter A-rich region. However, a tendency to increase in size was not observed for the oligo(A) sequences present in the Ljubljana strains. Strain Ljubljana II was isolated from the blood of a diseased laboratory worker, who acquired a laboratory infection with strain Ljubljana I (Avsic-Zupanc *et al.,* 1995). Ljubljana II was subsequently passaged three times in baby mice and then grown in BHK cells prior to our analysis. The analysis of the 3′-NCR of Ljubljana II revealed no sequence differences, including the oligo(A) structure, compared to Ljubljana I.

It is unknown whether the acquisition of a poly(A) structure, insertions, or deletions within the 3′-NCR offer any selective advantage that would favor the evolution of a particular length under certain conditions of virus growth. The absence of any correlation between strain-specific parameters, such as the year, place or source of isolation, and the 3′-NCR length, seems to argue against a specific selective mechanism. It rather suggests that deletions or insertions in the variable section of the 3′-NCR occur spontaneously and by chance. As outlined above the data suggest that the deletions were evolutionarily more recent events than the insertions. In fact, one might speculate that these deletions occurred

during propagation of the virus in the laboratory. The fact that those strains of TBE virus with the longest 3′-NCRs have been passaged only a limited number of times (compare Table 1) is consistent with this hypothesis. However, so far we have not observed such a deletion event during growth of TBE virus strain Neudoerfl under laboratory conditions. This question, however, will have to be investigated more extensively in the future.

Size variations of 3′-NCR regions among closely related strains have also recently been observed for other positive-stranded RNA viruses. Among alphaviruses, an avirulent variant of Semliki Forest virus exhibited 334 additional nucleotides compared to wild-type virus, which in part consisted of tandem repeat structures (Santagati *et al.,* 1994). The 3′-NCR region of pestiviruses can also be divided into a 3′-terminal conserved and a variable region. One strain of the pestivirus bovine viral diarrhea virus (strain Osloss) was found to have a 41-nucleotide-long deletion in the variable region of the 3′-NCR compared to two other strains of this virus (Deng and Brock, 1993). It is not known whether the short and variable 3′-NCR regions currently known for hepatitis C viruses actually represent the genomic 3′-termini (van Doorn, 1994). Considering the presence of internal poly(A) tracts in some of the TBE virus strains it is tempting to speculate that the poly(A) and poly(U) structures of hepatitis C virus might also be internal rather than representing the 3′-genomic termini. Heterogeneity within part of the 3′-NCR region may even be a widespread phenomenon among Flaviviridae and perhaps also other positive-stranded RNA virus families. It will be an interesting future question whether these differences affect the viral biology and pathogenicity as suggested by the above-mentioned avirulent variant of Semliki Forest virus that contains an insertion in the 3′-NCR (Santagati *et al.,* 1994). It should also be noted that with TBE virus, the long 3′-NCR strains Neudoerfl and particularily 263 exhibit a less virulent phenotype for mice than the short 3′-NCR strain Hypr (Kopezký *et al.,* 1991; unpublished observation). On the other hand, the Ljubljana strains which are distinguished from strains Neudoerfl and 263 basically only by their shorter homoadenosine box caused severe human disease (Avšič-Županc *et al.,* 1995). Genomic size differences of about 400 nucleotides as observed for the 3′-NCR of TBE virus illustrate a considerable flexibility of genomic packaging. Taking advantage of this flexibility and using the infectious cDNA technology now available for a growing number of flaviviruses it will be possible to investigate the role of the 3′-NCR for the biology of flaviviruses and its usefulness for modulating their pathogenic and immunogenic properties.

ACKNOWLEDGMENTS

The authors thank Melby Wilfinger, Angela Dohnal, and Walter Holzer for their excellent technical assistance.

REFERENCES

Avni, D., Shama, S., Loreni, F., and Meyuhas, O. (1994). Vertebrate mRNAs with a 5′-terminal pyrimidine tract are candidates for translational repression in quiescent cells: Characterization of the translational cis-regulatory element. *Mol. Cell. Biol.* **14,** 3822–3833.

Avšič-Županc, T., Poljak, M., Matičič, M., Radšel-Medveščck, A., LeDuc, J. W., Stiasny, K., Kunz, C., and Heinz, F. X. (1995). Laboratory acquired tick-borne meningoencephalitis: Characterisation of virus strains. *Clin. Diagn. Virol.* **4,** 51–59.

Behe, M. J. (1995). An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes. *Nucleic Acids Res.* **23,** 689–695.

Brinton, M. A., Fernandez, A. V., and Dispoto, J. H. (1986). The 3′-nucleotides of flavivirus genomic RNA form a conserved secondary structure. *Virology* **153,** 113–121.

Calisher, C. H., Karabatsos, N., Dalrymple, J. M., Shope, R. E., Porterfield, J. S., Westaway, E. G., and Brandt, W. E. (1989). Antigenic relationships between flaviviruses as determined by cross-neutralization tests with polyclonal antisera. *J. Gen. Virol.* **70,** 37–43.

Czyzykrzeska, M. F., Dominski, Z., Kole, R., and Millhorn, D. E. (1994). Hypoxia stimulates binding of a cytoplasmic protein to a pyrimidine-rich sequence in the 3′-untranslated region of rat tyrosine hydroxylase messenger RNA. *J. Biol. Chem.* **269,** 9940–9945.

Deng, R., and Brock, K. V. (1993). 5′ and 3′ untranslated regions of pestivirus genome: Primary and secondary structure analyses. *Nucleic Acids Res.* **8,** 1949–1957.

Dominguez, G., Wang, C.-Y., and Frey, T. K. (1990). Sequence of the genome RNA of rubella virus: Evidence for genetic rearrangement during togavirus evolution. *Virology* **177,** 225–238.

Hahn, C. S., Hahn, Y. S., Rice, C. M., Lee, E., Dalgarno, L., Strauss, E. G., and Strauss, J. H. (1987). Conserved elements in the 3′ untranslated region of flavivirus RNAs and potential cyclization sequences. *J. Mol. Biol.* **198,** 33–41.

Heinz, F. X. (1986). Epitope mapping of flavivirus glycoproteins. *Adv. Virus. Res.* **31,** 103–186.

Higgins, D. G., and Sharp, P. M. (1988). CLUSTAL: A package for performing multiple sequence alignments on a microcomputer. *Gene* **73,** 237–244.

Holzmann, H., Vorobyova, M. S., Ladyzhenskaya, I. P., Ferenczi, E., Kundi, M., Kunz, C., and Heinz, F. X. (1992). Molecular epidemiology of tick-borne encephalitis virus: Cross-protection between European and Far Eastern subtypes. *Vaccine* **10,** 345–349.

Hungnes, O., Tjotta, E., and Grinde, B. (1992). Mutations in the central polypurine tract of HIV-1 result in delayed replication. *Virology* **190,** 440–442.

Kopezký, J., Krivanec, K., and Tomková, E. (1991). Attenuated temperature-sensitive mutants of tick-borne encephalitis (TBE) virus isolated from natural focus. *In* "Modern Acarology Academia Prague and SPB" (F. Dusbábek and V. Bukva, Eds.), Vol. 2, pp. 11–19. Academic Publishing bv., The Hague.

Mandl, C. W., Heinz, F. X., and Kunz, C. (1988). Sequence of the structural proteins of tick-borne encephalitis virus (Western subtype) and comparative analysis with other flaviviruses. *Virology* **166,** 197–205.

Mandl, C. W., Heinz, F. X., Puchhammer-Stöckl, E., and Kunz, C. (1991a). Sequencing the termini of capped viral RNA by 5′–3′ ligation and PCR. *BioTechniques* **10,** 485–486.

Mandl, C. W., Heinz, F. X., Stöckl, E., and Kunz, C. (1989). Genome sequence of tick-borne encephalitis virus (Western subtype) and comparative analysis of nonstructural proteins with other flaviviruses. *Virology* **173,** 291–301.

Mandl, C. W., Holzmann, H., Kunz, C., and Heinz, F. X. (1993). Complete genomic sequence of Powassan virus: Evaluation of genetic elements in tick-borne versus mosquito-borne flaviviruses. *Virology* **194,** 173–184.

Mandl, C. W., Kunz, C., and Heinz, F. X. (1991b). Presence of poly(A) in a flavivirus: Significant differences between the 3′ noncoding re-

gions of the genomic RNAs of tick-borne encephalitis virus strains. *J. Virol.* **65,** 4070–4077.

Meyers, G., Rümenapf, T., and Thiel, H.-J. (1989). Ubiquitin in a togavirus. *Nature (London)* **341,** 491.

Pletnev, A. G., Yamshchikov, V. F., and Blinov, V. M. (1990). Nucleotide sequence of the genome and complete amino acid sequence of the polyprotein of tick-borne encephalitis virus. *Virology* **174,** 250–263.

Rice, C. M., Lenches, E. M., Eddy, S. R., Shin, S. J., Sheets, R. L., and Strauss, J. H. (1985). Nucleotide sequence of yellow fever virus: Implications for flavivirus gene expression and evolution. *Science* **229,** 726–733.

Santagati, M. G., Itäranta, P. V., Koskimies, P. R., Määttä, J. A., Salmi, A. A., and Hinkkanen, A. E. (1994). Multiple repeating motifs are found in the 3′-terminal nontranslated region of Semliki Forest virus A7 variant genome. *J. Gen. Virol.* **75,** 1499–1504.

Shamanin, V. A., Pletnev, A. G., Rubin, S. G., and Zlobin, V. I. (1990). Differentiation of strains of tick-borne encephalitis virus by means of RNA–DNA hybridization. *J. Gen. Virol.* **71,** 1505–1515.

Strauss, E. G., and Strauss, J. H. (1986). Structure and replication of the alphavirus genome. *In* "The Togaviridae and Flaviviridae" (S. Schlesinger and M. J. Schlesinger, Eds.), pp. 35-90. Plenum, New York.

van Doorn, L.-J. (1994). Molecular biology of the hepatitis C virus. *J. Med. Virol.* **43,** 345–356.

Wengler, G., Bradley, D. W., Collett, M. S., Heinz, F. X., Schlesinger, R. W., and Strauss, J. H. (1995). Flaviviridae. *In* "Virus Taxonomy. Classification and Nomenclature of Viruses" (F. A. Murphy, C. M. Fauquet, D. H. L. Bishop, S. A. Ghabrial, A. W. Jarvis, G. P. Martelli, M. A. Mayo, and M. D. Summers, Eds.), pp. 415–427. Springer-Verlag, Vienna/New York.

Wengler, G., and Castle, E. (1986). Analysis of structural properties which possibly are characteristics for the 3′-terminal sequences of the genome RNA of flaviviruses. *J. Gen. Virol.* **67,** 1183–1188.

Westaway, E. G., Brinton, M. A., Gaidamovitch, S. Y., Horzinek, M. C., Igarashi, A., Kaariainen, L., Lvov, D. K., Porterfield, J. S., Russell, P. K., and Trent, D. W. (1985). Flaviviridae. *Intervirology* **24,** 183–192.